2012 Presidential Candidates Donation Trend and the Correlation

Apoorv Mehrotra
SJSU ID: 014597635
MS in Software Engineering
San Jose State University

Bhavana Bangalore Sathyaprakash
SJSU ID: 014597245
MS in Software Engineering
San Jose State University

Namratha Bhat
SJSU ID: 014597258
MS in Software Engineering
San Jose State University

Pranjay Sagar
SJSU ID: 014611922
MS in Software Engineering
San Jose State University

Abstract

Data for fund collection tracking has been punched in by the Federal High Commission. It tracks the funds for the three Presidential Election Candidates for 2012. The project aims at identifying trends and correlations between donors and candidates.

Problem Statement

Federal Election Commission of the United States of America shared the complete details of the funding that was raised for the 2012 Presidential Elections. There were funds donated by people from all the states of the US. It needs to be analyzed to find out the correlation between the funds received for all the candidates.

Introduction

The idea behind the implementation of this project is to get familiar with the machine learning implementation. A lot of techniques have been used to unfold the funding campaign collection for the 2012 Presidential Elections. There are three candidates in total – namely Buddy Roemer, Gary Johnson, and Jill Stein. All of them submitted a report of funding received from numerous people all across the US. The track records maintained in the excel sheets include the details of the donors and the funds contributed by them for all the candidates contesting elections. This is being maintained by the Federal High Commission. They have used the received data to tabulate it and review it before paying the candidates. This becomes the data set for tracking the funding trends and also to find a correlation. In diving into the details of the excel per candidate basis, the grouping criteria is selected. There are certain common parameters that can be chosen for catching the trend of the funds made towards the election candidates. To categorize, it includes Occupation and State. Each parameter can be plotted for all the candidates to identify and display the actual aspects and funds accumulated for all the candidates.

Implementation

The entire data model has been compiled and built to find the trends of the funds that flowed in the 2012 Presidential Elections. The implementation has been done using the below listed technology stack:

1. IBM Watson
   Deep learning project delivered by IBM, termed after founder and first CEO of IBM, Thomas J. Watson, IBM Watson. It is a framework used to formulate the theoretical idea into its practical realization. This is used as a platform wherein the complete set of data was loaded. It provides support of a lot of utilities by handling the data itself and providing default access to the code

required. Using the key code and utilities generated by the IBM Watson, the further analysis of the code has been implemented.

2. Python
   The entire processing of the data has been implemented using the libraries and codes of python. The dataset shared with us has been cleaned, by aggregating multiple entries of the same donor. The second level of cleaning was done by aggregating all the similar occupations under the same category. The next stage that was implemented comprised of sorting the occupation data as per the number of donations received. Multiple libraries have been used and referred to which are illustrated below:

   a. pandas
      The tasks of data analysis and cleaning has been implemented using the pandas utility. It is used to make the data ready for the further implementation of graphs to track the flow of funds.

   b. matplot
      It is a plotting library though which all the graphs have been plotted.

   c. seaborn
      The library has been used for statistical data implementation. The major correlation has been judged through the stats and approaches represented by it.

Goals

The goal of the project is to take the available excel datasheets through the Federal High Commission and implement a data analysis to identify trends and correlation factors on it. The model has mainly three aspects.

1. Clean the data set given as per the identified categories.
2. Plot the graphs of various grouped categories to display the trend of funds,
3. Identify the correlation of funds among the three Presidential election candidates.

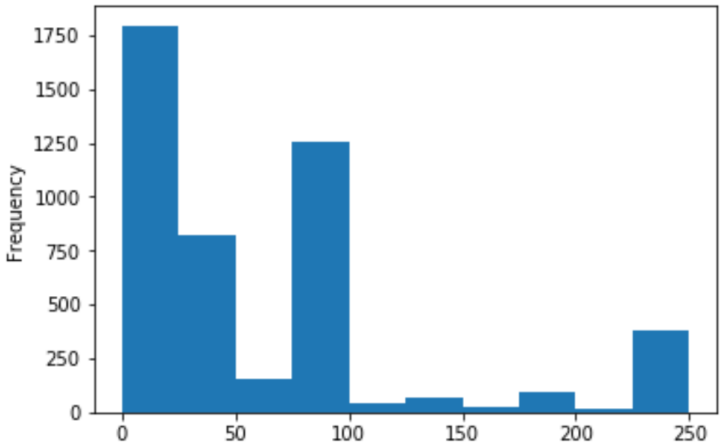
Challenges

The challenges that were faced during the implementation of this project are as below.

1. Data cleaning for occupation attribute was a major challenge since it required the identification of the categories that were to be aggregated together. The elimination part was to be taken care of to ensure the mitigation of data losing risk. Within the occupation column itself, multiple naming formats were followed for the same occupation. It also included spelling errors. Merging it to the centralized category was time and effort consuming.
2. Columns of different excel had different naming formats. Identifying it, filtering and merging also took some amount of time.
3. There were multiple factors of finding correlation. Prioritization of category for fund flow trend identification and correlation was identified at a later stage.

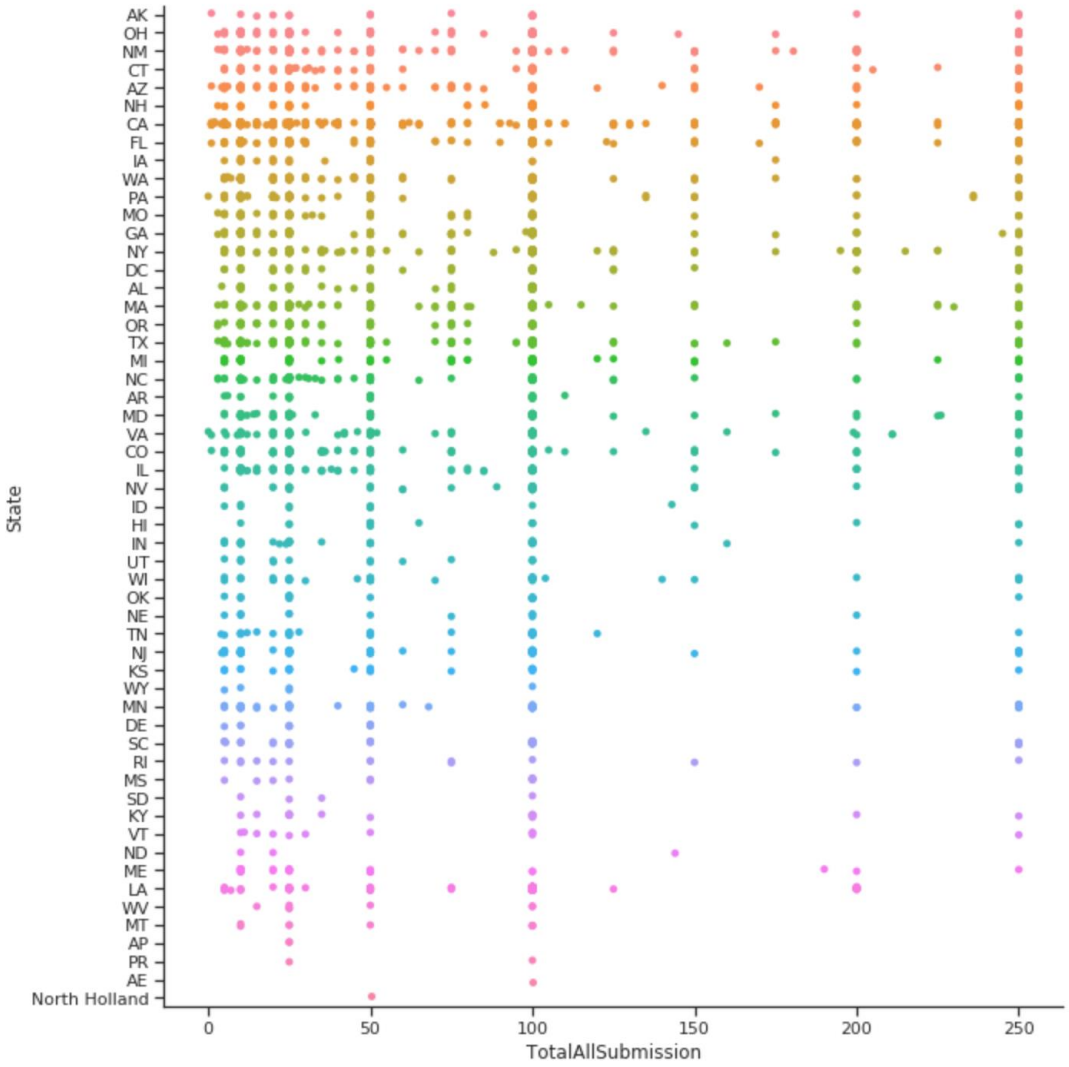
Conclusion

On analyzing the graphs plotted after the code data cleaning, grouping and sorting for all the three Presidential Candidates, below are the conclusions



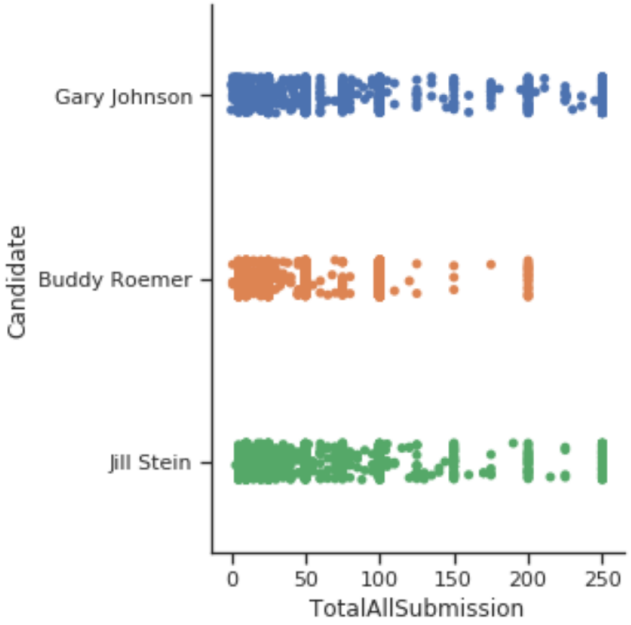
Histogram for the Amount donated for campaign vs Number of people donating the amount across entire campaign

Major donations were made in the range of $1 to $25. It was also noticed that contributions mainly confined within $100.

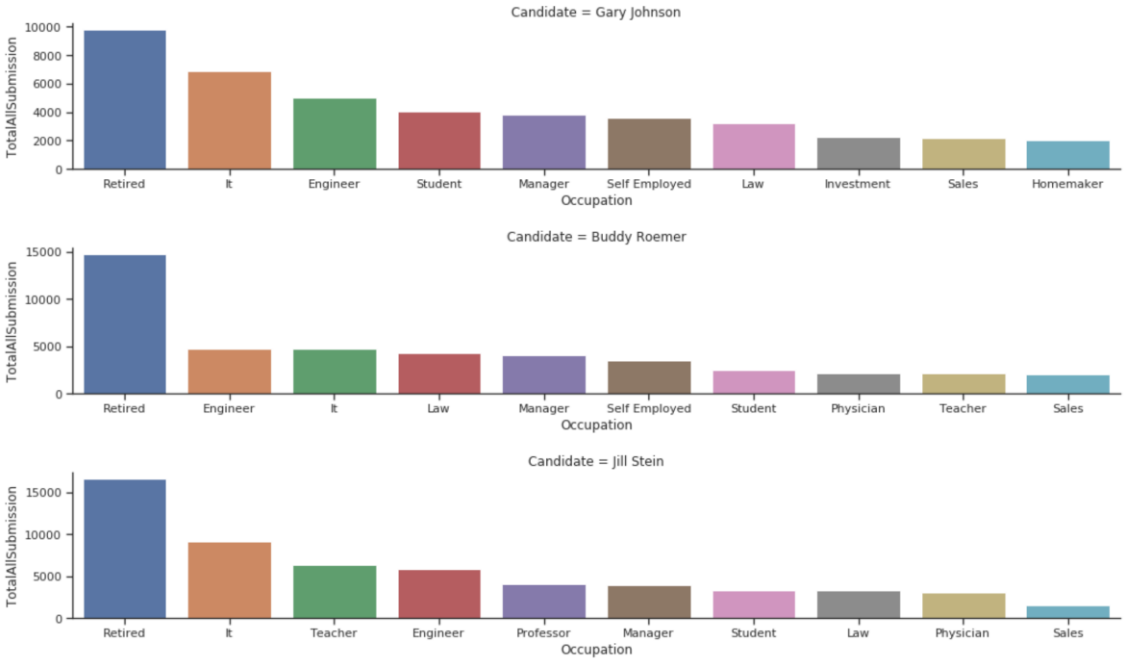


Top portion of the graph represents the states that made major contributions and the lower portion of the graph highlights the states that made lesser contribution towards the funds. Reading a few entries, California is observed as one of the major contributors, on the other hand Puerto Rico has minimal contribution.

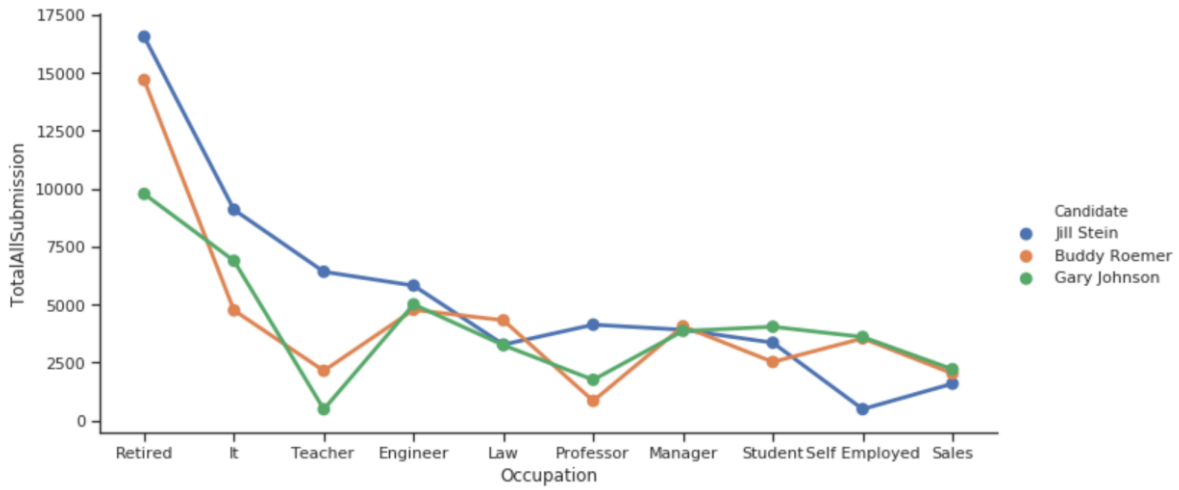
Next observation induced was on the basis of state-wise donation. In Washington, the pattern of donation followed is Jill Stein > Buddy Roemer > Gary Johnson. Whereas, in Florida all the three candidates received nearly equal donations.



From this graph, it was deduced that Jill Stein and Gary Johnson received more donations of higher denominations than Buddy Roemer.



From the above graph, top 10 occupations for each candidate are identified. It is noticed that most of the homemakers support Gary Johnson.



It is observed that the top three occupation that contributed towards the funds are Retired, IT and Engineer. On deep diving into the top 10 occupations, it is observed that retired people preferred Buddy Roemer and Jill Stein over Gary Johnson. Engineers, managers and Law practitioners tend to support all the candidates almost equally. Educators had an inclination towards Jill Stein.

References

https://transition.fec.gov/finance/2012matching/2012matching.shtml